Deep Hierarchy in Bandits

Joey Hong [1] **Branislav Kveton** [2] **Sumeet Katariya** [2] **Manzil Zaheer** [3] **Mohammad Ghavamzadeh** [4]

Abstract

Mean rewards of actions are often correlated. The form of these correlations may be complex and unknown a priori, such as the preferences of users for recommended products and their categories. To maximize statistical efficiency, it is important to leverage these correlations when learning. We formulate a bandit variant of this problem where the correlations of mean action rewards are represented by a *hierarchical Bayesian model* with latent variables. Since the hierarchy can have multiple layers, we call it *deep*. We propose a hierarchical Thompson sampling algorithm (`HierTS`) for this problem and show how to implement it efficiently for Gaussian hierarchies. The efficient implementation is possible due to a novel exact hierarchical representation of the posterior, which itself is of independent interest. We use this exact posterior to analyze the Bayes regret of `HierTS`. Our regret bounds reflect the structure of the problem, that the regret decreases with more informative priors, and can be recast to highlight reduced dependence on the number of actions. We confirm these theoretical findings empirically, in both synthetic and real-world experiments.

1. Introduction

A *contextual bandit* (Li et al., 2010; Chu et al., 2011) is a sequential decision-making problem where a *learning agent* sequentially interacts with an environment over n rounds. In each round, the agent observes a *context*, chooses one of K possible *actions*, and then receives a *reward* for the taken action. The agent aims to maximize its expected cumulative reward over n rounds. It does not know the mean rewards of the actions *a priori* and learns them by taking the actions. This forces the agent to choose between *exploring* actions to learn about them and *exploiting* the action with the highest estimated reward. As an example, in online shopping, the context can be a user's query, the actions are recommended products, and the reward is the indicator of a purchase (Yue & Guestrin, 2011; Kveton et al., 2015; Combes et al., 2015; Zong et al., 2016; Li et al., 2016).

In many practical problems, the action space is large and cannot be explored naively. However, the mean rewards of actions are correlated due to some underlying structure. As a result, exploration of one action can teach the agent about other actions, depending on how correlated they are, and this increases statistical efficiency of exploration. As an example, in online shopping, many products are semantically similar, and their relationships can be captured by a graphical model: both a keyboard and monitor are computer accessories; and both computer accessories and home theatre systems are electronic devices. Another example is classification with a bandit feedback, where the labels are clustered. For instance, the car and truck are vehicles, while the monkey and tiger are animals. We experiment with this kind of problems in Section 6.2.

Unfortunately, the aforementioned structure is not easy to represent in traditional bandit algorithms (Auer et al., 2002; Chapelle & Li, 2012; Kawale et al., 2015; Sen et al., 2017). As an example, *Thompson sampling (TS)* (Thompson, 1933; Chapelle & Li, 2012; Agrawal & Goyal, 2012; Russo & Van Roy, 2014) is a natural and elegant strategy for exploration in bandit problems. However, in structured action spaces, TS would need to model correlations between all actions using their joint posterior. This could be computationally costly in practice, as observing the reward of a single action may require recomputing the joint posterior of all actions. Therefore, it is not immediately obvious how to design an efficient TS algorithm when the mean rewards of actions are correlated.

In this work, we study a structured bandit problem where the action space has a *hierarchical structure*. This structure is represented using a *hierarchical Bayesian model* (Lindley & Smith, 1972; Zhang & Yang, 2017), where each action is a leaf node in a tree. Each node is associated with a *node parameter*, which is drawn i.i.d. from a distribution parameterized by its parent's parameter. This hierarchical structure not only models many practical bandit tasks, such as online shopping, but admits an efficient exploration because the

[1] University of California, Berkeley [2] Amazon [3] DeepMind [4] Google Research. Correspondence to: Joey Hong <joey_hong@berkeley.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

joint posterior can be factorized and efficiently updated.

We make the following contributions. First, we formalize a hierarchical Bayesian model $\mathcal{T}$ of our environment. Second, we propose a *Thompson sampling* algorithm `HierTS`. The main novelty in the design of `HierTS` is that the posterior factors along $\mathcal{T}$, which permits *exact posterior sampling* and *computationally-efficient updates*. The posterior has a closed form in multi-armed and contextual linear bandits with Gaussian rewards. We derive Bayes regret bounds for `HierTS` that capture the structure of $\mathcal{T}$ and the impact of priors. The bounds show increased statistical efficiency due to the hierarchy and can improve upon classical results in non-constant factors. We validate these theoretical results empirically and also apply `HierTS` to a real-world classification problem with label hierarchy.

2. Setting

We use the following notation. Random variables are capitalized. For any positive integer n, we denote by $[n]$ the set $\{1, \dots, n\}$. We let $\mathbb{1}\{\cdot\}$ be the indicator function. The i-th entry of vector v is v_i. If the vector is already indexed, such as v_j, we write $v_{j,i}$. We use $\tilde{O}$ for the big O notation up to logarithmic factors.

We consider a learning agent that interacts with a contextual bandit over n rounds (Li et al., 2010; Chu et al., 2011). In round $t \in [n]$, the agent observes *context* $X_t \in \mathcal{X} \subseteq \mathbb{R}^d$, takes an *action* A_t from an *action set* $\mathcal{A}$ of size K, and then observes a *stochastic reward* $Y_t = r(X_t, A_t) + \varepsilon_t$, where $r : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ is a *reward function* and ε_t is independent σ^2-sub-Gaussian noise.

Our problem is structured. In particular, the action set $\mathcal{A}$ progressively breaks into finer clusters of actions with similar rewards. This decomposition is represented by a tree $\mathcal{T}$ (Figure 1) over nodes $\mathcal{V} \subset \mathbb{N}$. Without loss of generality, the root has index 1. Each *leaf* of $\mathcal{T}$ corresponds to an action $a \in \mathcal{A}$ and we call it an *action node*. Each *internal node* of $\mathcal{T}$ has at least two and at most b children, where b is the *branching factor*. The *height of the tree* is h and the *height of node* $i \in \mathcal{V}$ is $h_i \leq h$. The height of the leaves is 0 and the height of the root is h. For any node $i \in \mathcal{V}$, we denote its *parent* by $\mathsf{pa}(i)$ and its *children* by $\mathsf{ch}(i)$. An *ancestor* of node i is any node on a direct path from node i to the root, and node i is a *descendant* of any node on that path. With a slight abuse of notation, we use $\mathcal{A} \subseteq \mathcal{V}$ to refer to both the action set and leaves of $\mathcal{T}$, and sometimes index action nodes by a to stress their role.

The reward function is parameterized by *model parameters* $\Theta = (\theta_i)_{i \in \mathcal{V}}$, where θ_i is the parameter of node i. The *true model parameters* are $\Theta_* = (\theta_{*,i})_{i \in \mathcal{V}}$ and we assume that they are generated as

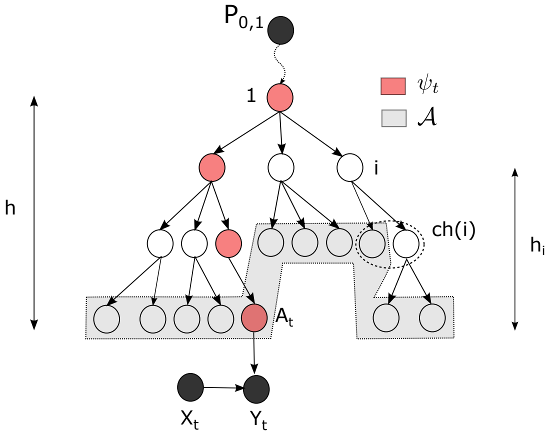

Figure 1. Graphical model of our environment. The drawing depicts our notation: children $\mathsf{ch}(i)$ of node i, action nodes $\mathcal{A}$, and updated nodes ψ_t after action A_t is taken. The height of the tree is $h = 3$ and that of node i is $h_i = 2$.

$$\begin{aligned} \theta_{*,1} &\sim P_{0,1}\,, && (1) \\ \theta_{*,i} \mid \theta_{*,\mathsf{pa}(i)} &\sim P_{0,i}(\cdot \mid \theta_{*,\mathsf{pa}(i)})\,, && \forall i \in \mathcal{V} \setminus \{1\}\,, \\ Y_t \mid X_t, \theta_{*,A_t} &\sim P(\cdot \mid X_t; \theta_{*,A_t})\,, && \forall t \in [n]\,. \end{aligned}$$

Here $P_{0,1}$ is the prior distribution of the root node, called a *hyper-prior*; $P_{0,i}(\cdot \mid \theta_{*,\mathsf{pa}(i)})$ is the *conditional prior distribution* of node i, parameterized by the sampled value of its parent $\theta_{*,\mathsf{pa}(i)}$; and $P(\cdot \mid x; \theta_{*,a})$ is the *reward distribution* of action a in context x. We use $r(x, a; \Theta)$ to denote the mean reward of action a in context x under model parameters Θ and define it as $r(x, a; \Theta) = \mathbb{E}_{Y \sim P(\cdot \mid x; \theta_a)}[Y]$. Thus $r(x, a; \Theta)$ depends only on one parameter in Θ. The generative process in (1) relates any two node parameters to each other, through the lowest common ancestor. This induces complex correlations that can be used for efficient exploration. We discuss motivating examples for this setting in Section 1.

The goal is to minimize the n-round regret defined as

$$\mathcal{R}(n; \Theta_*) = \mathbb{E}\left[\sum_{t=1}^{n} r(X_t, A_{t,*}; \Theta_*) - r(X_t, A_t; \Theta_*)\right],$$

where $A_{t,*} = \arg\max_{a \in \mathcal{A}} r(X_t, a; \Theta_*)$ is the optimal action in round t given context X_t. In this work, we assume that the parameters Θ_* are also random. We define the n-round *Bayes regret* as $\mathcal{BR}(n) = \mathbb{E}\left[\mathcal{R}(n; \Theta_*)\right]$, which takes an additional expectation over Θ_*. While weaker than the traditional frequentist regret $\mathcal{R}(n; \Theta_*)$, the Bayes regret is a practical performance measure, when the average performance across multiple instances of model parameters is of interest (Russo & Van Roy, 2014; Hong et al., 2020).

Algorithm 1 `HierTS`: Hierarchical Thompson sampling.

```
Input: Tree $\mathcal{T}$ with height $h$, all priors $P_{0,\cdot}$ in (1)
Initialize all posteriors $P_{1,\cdot} \leftarrow P_{0,\cdot}$
for $t = 1, \dots, n$ do
    Sample $\theta_{t,1} \sim P_{t,1}$
    for $\ell = h - 1, \dots, 0$ do
        for $i \in \mathcal{V}_\ell$ do
            Sample $\theta_{t,i} \sim P_{t,i}(\cdot \mid \theta_{t,\mathsf{pa}(i)})$
        end for
    end for
    $\Theta_t \leftarrow (\theta_{t,i})_{i \in \mathcal{V}}$
    Take action $A_t \leftarrow \arg\max_{a \in \mathcal{A}} r(X_t, a; \Theta_t)$
    Observe $Y_t \sim P(\cdot \mid X_t; \theta_{*,A_t})$
    Compute new posteriors $P_{t+1,\cdot}$
end for
```

3. Algorithm

Since our environment is a graphical model (Figure 1), we explore using *Thompson sampling (TS)* (Thompson, 1933; Chapelle & Li, 2012; Agrawal & Goyal, 2012; Russo & Van Roy, 2014). The main challenge in our algorithm design are *latent variables*. Specifically, the rewards of action nodes are observed and permit direct learning of their parameters, $\theta_{*,a}$ for $a \in \mathcal{A}$. In contrast, parameters $\theta_{*,i}$ of the internal nodes $i \in \mathcal{V} \setminus \mathcal{A}$ are only indirectly observed through their descendant action nodes, and thus are latent.

It is unclear if modeling of the latent variables is necessary. To discuss alternatives, we need to introduce some notation. Let $H_t = (X_\ell, A_\ell, Y_\ell)_{\ell \in [t-1]}$ be the *history* of all interactions of the agent until round t and $\Theta_{*,\mathcal{A}} = (\theta_{*,a})_{a \in \mathcal{A}}$ be the true model parameters of the action nodes. Because the mean rewards of actions depend only on $\Theta_{*,\mathcal{A}}$, the most natural solution to our problem is TS over a joint posterior $\Theta_{*,\mathcal{A}} \mid H_t$. This has two challenges. First, the exact posterior involves complex correlations, due to the dependencies in $\theta_{*,a}$ induced by the generative process in (1). These correlations remain when the latent variables are marginalized out, and may not allow computationally-efficient sampling from $\Theta_{*,\mathcal{A}} \mid H_t$. Second, the uncertainty of each $\theta_{*,a} \mid H_t$ could be modeled individually. While computationally efficient, this may not be sound and would not be statistically efficient. We propose *exact sampling* from $\Theta_{*,\mathcal{A}} \mid H_t$ that is both *computationally and statistically efficient*.

3.1. Hierarchical Sampling

Our approach is based on hierarchical sampling (Andrieu et al., 2003; Doucet et al., 2001), where the model parameters Θ are sampled similarly to the generative process in (1). To explain it, we introduce the following notation. For the root node, $P_{t,1}(\theta) = \mathbb{P}(\theta_{*,1} = \theta \mid H_t)$ denotes the posterior distribution of its parameter in round t, which we also call a *hyper-posterior*. For any other node i,

$$P_{t,i}(\theta \mid \theta_p) = \mathbb{P}\left(\theta_{*,i} = \theta \mid \theta_{*,\mathsf{pa}(i)} = \theta_p, H_{t,i}\right) \tag{2}$$

is the posterior distribution of its parameter conditioned on $\theta_{*,\mathsf{pa}(i)} = \theta_p$ in round t, where $H_{t,i}$ is a subset of interactions in H_t where A_ℓ is a descendant of node i. In (2), H_t could be replaced by $H_{t,i}$ because the posterior of $\theta_{*,i}$ is independent of the other observations given the value of the parent parameter θ_p. This structure is critical to the computational efficiency of our approach and is also used in the regret analysis (Section 5).

Assuming that all posteriors $P_{t,i}$ can be computed efficiently, it is trivial to propose a hierarchical Thompson sampling algorithm for our problem. We call it `HierTS` and present its pseudo-code in Algorithm 1. In round t, `HierTS` works as follows. First, we sample the root parameter $\theta_{t,1}$. After that, we iterate over all nodes and sample node parameters whose parents are already sampled. Specifically, we define $\mathcal{V}_\ell = \{i \in \mathcal{V} : h_i = \ell\}$ as the subset of nodes at height ℓ and then sample $\theta_{t,i}$ for $i \in \mathcal{V}_\ell$, starting from the children of the root at height $\ell = h - 1$ all the way to the leaves at height $\ell = 0$. By design, $\Theta_t = (\theta_{t,i})_{i \in \mathcal{V}}$ is a valid posterior sample, generated hierarchically. Finally, `HierTS` takes an optimistic action with respect to Θ_t, observes Y_t, and then updates its posterior.

Note that `HierTS` samples parameters at all action nodes. It is possible to leverage the tree structure to prune sub-trees with actions that are unlikely to have high mean rewards. For example, Sen et al. (2021) propose beam search over a tree to only evaluate a subset of actions. Such computational improvements can be easily incorporated into `HierTS`. We view them as orthogonal to our main contribution, which is a statistically-efficient exploration using the tree structure.

3.2. Efficient Posterior Computation

The main technical novelty in `HierTS` is that the posteriors $P_{t,i}$ can be maintained efficiently. We show it as follows.

Fix any node i, its value θ, and the value of its parent θ_p. By Bayes rule, we have

$$P_{t,i}(\theta \mid \theta_p) \propto \mathcal{L}_{t,i}(\theta) P_{0,i}(\theta \mid \theta_p)\,, \tag{3}$$

where $\mathcal{L}_{t,i}(\theta) = \mathbb{P}(H_{t,i} \mid \theta_{*,i} = \theta)$ is the likelihood of observations $H_{t,i}$ whose ancestor is node i with value θ. Note that $\mathcal{L}_{t,i}(\theta)$ can be further decomposed as

$$\mathcal{L}_{t,i}(\theta) = \textstyle\prod_{j \in \mathsf{ch}(i)} \tilde{\mathcal{L}}_{t,j}(\theta)\,, \tag{4}$$

where $\tilde{\mathcal{L}}_{t,j}(\theta) = \mathbb{P}\left(H_{t,j} \mid \theta_{*,\mathsf{pa}(j)} = \theta\right)$ is the likelihood of observations $H_{t,j}$ whose ancestor is child node j and the value of its parent is θ. This identity follows from two facts. First, $\theta_{*,j}$ are conditionally independent of each other given

Algorithm 2 Updated statistics in round t. The dot notation means that the likelihoods are updated for all parameter values, which is possible in Gaussian models (Section 4).

```
Initialize $\mathcal{L}_{t+1,\cdot} \leftarrow \mathcal{L}_{t,\cdot}$
$i \leftarrow A_t$
$\mathcal{L}_{t+1,i}(\cdot) \leftarrow P(Y_t \mid X_t; \cdot)\mathcal{L}_{t,i}(\cdot)$
$\tilde{\mathcal{L}}_{t+1,i}(\cdot) \leftarrow \int_\theta \mathcal{L}_{t+1,i}(\theta) P_{0,i}(\theta \mid \cdot) \, \mathrm{d}\theta$
repeat
  $i \leftarrow \mathsf{pa}(i)$
  $\mathcal{L}_{t+1,i}(\cdot) \leftarrow \prod_{j \in \mathsf{ch}(i)} \tilde{\mathcal{L}}_{t+1,j}(\cdot)$
  if $i > 1$ then
    $\tilde{\mathcal{L}}_{t+1,i}(\cdot) \leftarrow \int_\theta \mathcal{L}_{t+1,i}(\theta) P_{0,i}(\theta \mid \cdot) \, \mathrm{d}\theta$
  end if
until $i = 1$
```

$\theta_{*,i} = \theta$. Second, $H_{t,j}$ is independent of $\theta_{*,i}$ given $\theta_{*,j}$. At a high level, each $\tilde{\mathcal{L}}_{t,j}(\theta)$ can be viewed as the likelihood of an aggregate observation at node j, from all leaves that descend from node j, under the assumption that $\theta_{*,i} = \theta$ (Section 4.1).

Finally, each $\tilde{\mathcal{L}}_{t,j}(\theta)$ can be computed as

$$\tilde{\mathcal{L}}_{t,j}(\theta) = \int_{\theta'} \mathcal{L}_{t,j}(\theta') P_{0,j}(\theta' \mid \theta) \, \mathrm{d}\theta' \,, \tag{5}$$

where $\mathcal{L}_{t,j}(\theta')$ is the likelihood of observations $H_{t,j}$ whose ancestor is node j with value θ'. Note that $\mathcal{L}_{t,j}(\theta')$ can be further rewritten as in (4), which gives rise to our recursive computation of the posterior.

The pseudo-code for updating $\mathcal{L}_{t,i}$ and $\tilde{\mathcal{L}}_{t,i}$ after round t is shown in Algorithm 2. After this, (3) has to be recomputed for all nodes i on the path from A_t to the root. In general, (5) is hard to compute due to the integral over θ'. However, in Gaussian graphical models (Section 4), this can be done in a closed form. In practice, (5) can be approximated for arbitrary distributions using approximate inference, either variational or MCMC (Doucet et al., 2001).

4. Gaussian Hierarchy

In this section, we instantiate the environment in (1) as a hierarchical Gaussian model (Koller & Friedman, 2009) and derive its posterior. The model is defined as

$$\begin{aligned} \theta_{*,1} &\sim \mathcal{N}(\mu_1, \Sigma_{0,1}) \,, \\ \theta_{*,i} \mid \theta_{*,\mathsf{pa}(i)} &\sim \mathcal{N}(\theta_{*,\mathsf{pa}(i)}, \Sigma_{0,i}) \,, \qquad \forall i \in \mathcal{V} \setminus \{1\} \,, \\ Y_t \mid X_t, \theta_{*,A_t} &\sim \mathcal{N}(X_t^\top \theta_{*,A_t}, \sigma^2) \,, \qquad \forall t \in [n] \,, \end{aligned} \tag{6}$$

where $\theta_{*,i} \in \mathbb{R}^d$ are the node parameters and $\Sigma_{0,i}$ are the covariance matrices that control the closeness of $\theta_{*,i}$ and $\theta_{*,\mathsf{pa}(i)}$. The mean reward is defined as $r(x, a; \Theta) = x^\top \theta_a$. The hierarchical structure is motivated by multi-label classification (Prabhu et al., 2018; Yu et al., 2020a), where X_t is a feature vector, A_t is its predicted label, and Y_t indicates if the label is correct. We return to this application in Section 6.2. When $d = 1$ and $X_t = 1$, we recover a K-armed Gaussian bandit, where $\theta_{*,a}$ is the mean reward of action a. We assume that the agent knows the hyper-prior mean μ_1, all covariances $\Sigma_{0,i}$, and reward noise σ. This is only used in the regret analysis, where we study exact posterior sampling. In our experiments (Section 6.2), we learn these quantities from data.

The special case of multi-armed bandits also shows computational gains over naive posterior sampling. Specifically, due to the dependencies in (1), $\Theta_{*,\mathcal{A}} \mid H_t$ is a multivariate Gaussian with K dimensions. To sample from it, we can compute the the root of the posterior covariance and multiply it by a K-dimensional standard normal vector, which takes $O(K^3)$ time. In contrast, sampling in `HierTS` takes $O(|\mathcal{V}|)$ time. When each internal node of $\mathcal{T}$ has at least 2 children, which is without loss of generality, $|\mathcal{V}| \leq 2K$ and our computational gain is $O(K^2)$. Now we present closed-form posteriors for hierarchies of Gaussian and contextual linear models.

4.1. Multi-Armed Bandit

We start with a K-armed Gaussian bandit. In this setting, each node $i \in \mathcal{V}$ is associated with a single scalar parameter $\theta_{*,i} \in \mathbb{R}$, and its initial uncertainty is described by conditional prior variance $\Sigma_{0,i} = \sigma_{0,i}^2 \in \mathbb{R}$. The posteriors for this model are derived in Appendix A.1 and stated below.

For any node i, the posterior of $\theta_{*,i}$ given $\theta_{*,\mathsf{pa}(i)} = \theta_p$ is $P_{t,i}(\theta \mid \theta_p) = \mathcal{N}(\theta; \hat{\theta}_{t,i}, \hat{\sigma}_{t,i}^2)$. If node i is an internal node,

$$\begin{aligned} \hat{\sigma}_{t,i}^{-2} &= \sigma_{0,i}^{-2} + \sum_{j \in \mathsf{ch}(i)} \tilde{\sigma}_{t,j}^{-2} \,, \\ \hat{\theta}_{t,i} &= \hat{\sigma}_{t,i}^2 \left(\sigma_{0,i}^{-2} \theta_p + \sum_{j \in \mathsf{ch}(i)} \tilde{\sigma}_{t,j}^{-2} \tilde{\theta}_{t,j} \right) . \end{aligned} \tag{7}$$

At the root ($i = 1$), we set $\theta_p = \mu_1$. The child parameters $\tilde{\theta}_{t,j}$ and $\tilde{\sigma}_{t,j}$ are computed recursively as follows. If node j is an action node, then

$$\begin{aligned} \tilde{\sigma}_{t,j}^2 &= \sigma_{0,j}^2 + \frac{\sigma^2}{|\mathcal{S}_{t,j}|} \,, \\ \tilde{\theta}_{t,j} &= \frac{1}{|\mathcal{S}_{t,j}|} \sum_{\ell \in \mathcal{S}_{t,j}} Y_\ell \,, \end{aligned} \tag{8}$$

where $\mathcal{S}_{t,j} = \{\ell < t : A_\ell = j\}$ are the rounds where action j is taken before round t. If node j is an internal node,

$$\begin{aligned} \tilde{\sigma}_{t,j}^2 &= \sigma_{0,j}^2 + M^{-1} \,, \\ \tilde{\theta}_{t,j} &= M^{-1} \sum_{k \in \mathsf{ch}(j)} \tilde{\sigma}_{t,k}^{-2} \tilde{\theta}_{t,k} \,, \end{aligned} \tag{9}$$

where $M = \sum_{k \in \mathsf{ch}(j)} \tilde{\sigma}_{t,k}^{-2}$. The new child parameters $\tilde{\theta}_{t,k}$ and $\tilde{\sigma}_{t,k}$ are computed recursively, using either (8) or (9). Finally, if node i is action node, its posterior has a standard form of

$$\hat{\sigma}_{t,i}^{-2} = \sigma_{0,i}^{-2} + \sigma^{-2} \left|\mathcal{S}_{t,i}\right| \,,$$
$$\hat{\theta}_{t,i} = \hat{\sigma}_{t,i}^{2} \left(\sigma_{0,i}^{-2} \theta_p + \sigma^{-2} \sum_{\ell \in \mathcal{S}_{t,i}} Y_\ell\right) .$$

The recursive update in (9) can be derived from the observation that $\tilde{\mathcal{L}}_{t,j}(\theta) \propto \exp\left[-\frac{1}{2}\tilde{\sigma}_{t,j}^{-2}(\theta - \tilde{\theta}_{t,j})^2\right]$ holds for any node j and the value of its parent θ. The closed-form of the posterior $P_{t,i}(\theta \mid \theta_p)$ is a direct combination of this result and the derivations in Section 3.2. The update in (9) also has an intuitive interpretation. Although we get Gaussian observations at action nodes, as in (8), they propagate to higher nodes in the tree through (9). These nodes then act as noisy observations of their parents with mean $\tilde{\theta}_{t,j}$ and variance $\tilde{\sigma}_{t,j}^2$. This allows us to overcome the problem of latent variables in our model. The posterior in (7) is just a function of higher-level observations in all children of node i. To have closed forms of these quantities, we rely heavily on the properties of Gaussian random variables.

4.2. Contextual Linear Bandit

We now consider the general case in (6). This model can be viewed as a hierarchy of linear models (Yue & Guestrin, 2011; Abbasi-Yadkori et al., 2011) indexed by actions. The posteriors for this model are derived in Appendix A.2 and stated below.

For any node i, the posterior of $\theta_{*,i}$ given $\theta_{*,\mathsf{pa}(i)} = \theta_p$ is $P_{t,i}(\theta \mid \theta_p) = \mathcal{N}(\theta; \hat{\theta}_{t,i}, \hat{\Sigma}_{t,i})$. If node i is an internal node,

$$\hat{\Sigma}_{t,i}^{-1} = \Sigma_{0,i}^{-1} + \sum_{j \in \mathsf{ch}(i)} \tilde{\Sigma}_{t,j}^{-1} \,, \qquad (10)$$
$$\hat{\theta}_{t,i} = \hat{\Sigma}_{t,i} \left(\Sigma_{0,i}^{-1} \theta_p + \sum_{j \in \mathsf{ch}(i)} \tilde{\Sigma}_{t,j}^{-1} \tilde{\theta}_{t,j}\right) .$$

At the root ($i = 1$), we set $\theta_p = \mu_1$. The child parameters $\tilde{\theta}_{t,j}$ and $\tilde{\Sigma}_{t,j}$ are computed recursively as follows. If node j is an action node, then

$$\tilde{\Sigma}_{t,j} = \Sigma_{0,j} + G_{t,j}^{-1} \,, \qquad (11)$$
$$\tilde{\theta}_{t,j} = \sigma^{-2} G_{t,j}^{-1} \sum_{\ell \in \mathcal{S}_{t,j}} X_\ell Y_\ell \,,$$

where $\mathcal{S}_{t,j} = \{\ell < t : A_\ell = j\}$ are the rounds where action j is taken before round t and $G_{t,j} = \sigma^{-2} \sum_{\ell \in \mathcal{S}_{t,j}} X_\ell^\top X_\ell$ is the outer product of the corresponding feature vectors. If node j is an internal node, then

$$\tilde{\Sigma}_{t,j} = \Sigma_{0,j} + M^{-1} \,, \qquad (12)$$
$$\tilde{\theta}_{t,j} = M^{-1} \sum_{k \in \mathsf{ch}(j)} \tilde{\Sigma}_{t,j}^{-1} \tilde{\theta}_{t,k} \,,$$

where $M = \sum_{k \in \mathsf{ch}(j)} \tilde{\Sigma}_{t,k}^{-1}$. The new child parameters $\tilde{\theta}_{t,k}$ and $\tilde{\Sigma}_{t,k}$ are computed recursively, depending on whether k is an action node or not. Finally, if node i is action node, its posterior has a standard form of

$$\hat{\Sigma}_{t,i}^{-1} = \Sigma_{0,i}^{-1} + G_{t,i} \,,$$
$$\hat{\theta}_{t,i} = \hat{\Sigma}_{t,i} \left(\Sigma_{0,i}^{-1} \theta_p + \sigma^{-2} \sum_{\ell \in \mathcal{S}_{t,i}} X_\ell Y_\ell\right) .$$

The recursive update in (12) can be derived from the observation that $\tilde{\mathcal{L}}_{t,j}(\theta) \propto \exp\left[-\frac{1}{2}(\theta - \tilde{\theta}_{t,j})^\top \tilde{\Sigma}_{t,j}^{-1} (\theta - \tilde{\theta}_{t,j})\right]$ for any node j and the value of its parent θ. The posterior in $P_{t,i}(\theta \mid \theta_p)$ is a direct combination of this result and the derivations in Section 3.2. As in Section 4.1, our recursive update can be viewed as propagation of observations from action nodes to higher nodes in the tree.

5. Analysis

This section is primarily devoted to the Gaussian bandit in Section 4.1. We present the key lemmas, the main result, and discuss them. All proofs are deferred to Appendix B. In Section 5.4, we state a regret bound for the contextual bandit in Section 4.2. Its proof is deferred to Appendix C.

5.1. Key Steps in the Analysis

We start with the observation that the hierarchical posterior sampling in Section 3.1 is just an efficient implementation of joint posterior sampling over the action node parameters $\Theta_{\mathcal{A}}$. Since our model is a *Gaussian graphical model*, this posterior is a multivariate Gaussian (Koller & Friedman, 2009). This is because any conditioning or marginalization does not change the model class. This observation allows us to prove the following lemma.

Lemma 1. *For any $\delta > 0$, the Bayes regret of* `HierTS` *is bounded as*

$$\mathcal{BR}(n) \leq \sqrt{2n\mathcal{G}(n)\log(1/\delta)} + \sqrt{2/\pi}\sigma_{\max} K n \delta \,,$$

where $\mathcal{G}(n) = \mathbb{E}\left[\sum_{t=1}^n \bar{\sigma}_{t,A_t}^2\right]$ is a statistical complexity term, *$\bar{\sigma}_{t,A_t}^2 = \text{var}\left[\theta_{t,A_t} \mid H_t\right]$ is the marginal posterior variance of the mean reward of action A_t in round t, and $\sigma_{\max}$ is the maximum marginal prior width at an action node.*

The second term in Lemma 1 is constant in n for $\delta = 1/n$. Therefore, we focus on the first $\tilde{O}(\sqrt{n})$ term. Also note that

$\bar{\sigma}^2_{t,A_t}$ in $\mathcal{G}(n)$ is none of the conditional posterior variances derived in Section 4.1. We show how to decompose it into these variances next.

To relate the marginal posterior variance, which bounds the expected regret, to conditional posterior variances, which represent our model uncertainty, we adopt the following update-centric notation. We denote the list of nodes from the root to the action node A_t in round t by ψ_t. The length of ψ_t is L_t. To illustrate the notation, $\psi_t(1) = 1$ is the root, $\psi_t(L_t) = A_t$ is the action node, and $\psi_t(L_t - 1) = \mathsf{pa}(A_t)$ is its parent. Figure 1 visualizes ψ_t. Now we are ready to relate the two quantities.

Lemma 2. *In any round t, the marginal posterior variance in action node A_t decomposes as*

$$\bar{\sigma}^2_{t,A_t} = \sum_{i=1}^{L_t} \left(\prod_{j=i+1}^{L_t} \frac{\hat{\sigma}^4_{t,\psi_t(j)}}{\sigma^4_{0,\psi_t(j)}} \right) \hat{\sigma}^2_{t,\psi_t(i)} \,.$$

The last piece is a lower bound, which shows that each term in Lemma 2 can be bounded by the posterior update of the corresponding node i, representing our information gain.

Lemma 3. *Fix any round t and $i \in [L_t]$. Then*

$$\hat{\sigma}^{-2}_{t+1,\psi_t(i)} - \hat{\sigma}^{-2}_{t,\psi_t(i)} \geq \sigma^{-2} c^{i - L_t} \left(\prod_{j=i+1}^{L_t} \frac{\hat{\sigma}^4_{t,\psi_t(j)}}{\sigma^4_{0,\psi_t(j)}} \right),$$

where $c = 1 + \sigma^2_{0,\max}/\sigma^2$ and $\sigma^2_{0,\max} = \max_{i \in \mathcal{V}} \sigma^2_{0,i}$ is the maximum prior variance. For $\sigma \geq \sigma_{0,\max}$, we have $c = 2$.

The constant c in Lemma 3 is small when the reward noise is higher than all prior widths in $\mathcal{T}$. This property can be always attained by initial forced exploration of all actions.

5.2. Regret Bound

Now we are ready to present our main result. Recall that h is the height of $\mathcal{T}$, h_i is the height of node i, and that the action nodes have height 0 (Section 2).

Theorem 4. *For any $\delta > 0$, the Bayes regret of* `HierTS` *is bounded as*

$$\mathcal{BR}(n) \leq \sqrt{2n\mathcal{G}(n)\log(1/\delta)} + \sqrt{2/\pi}\sigma_{\max} K n \delta \,,$$

where $\mathcal{G}(n) = \sum_{i \in \mathcal{V}} c^{h_i} w_i$ and c is a constant defined in Lemma 3. For an action node i, $h_i = 0$ and

$$w_i = \frac{\sigma^2_{0,i}}{\log\left(1 + \frac{\sigma^2_{0,i}}{\sigma^2}\right)} \log\left(1 + \frac{\sigma^2_{0,i} n}{\sigma^2}\right).$$

For an internal node i, $h_i > 0$ and

$$w_i = \frac{\sigma^2_{0,i}}{\log\left(1 + \frac{\sigma^2_{0,i}}{\sigma^2}\right)} \log\left(1 + \sigma^2_{0,i} \sum_{j \in \mathsf{ch}(i)} \sigma^{-2}_{0,j}\right).$$

When $\delta = 1/n$, the above bound is $\tilde{O}(\sqrt{n\,|\mathcal{V}|})$, where n is the horizon and $|\mathcal{V}|$ is the number of nodes, which is also the number of learned parameters. The dependence on the horizon n is standard. As $w_i = \tilde{O}(\sigma^2_{0,i})$, the contribution of each node i to the regret is proportional to its prior width. Thus the regret decreases when the model is more certain. One notable dependence in $\mathcal{G}(n)$ is exponential scaling with height c^{h_i}. This is not problematic, as the number of nodes with high h_i is exponentially smaller than those with lower h_i (Section 5.3).

Theorem 4 also recovers a well-known Bayes regret bound for K-armed bandits (Russo & Van Roy, 2014). The reason is that a K-armed bandit can be viewed as a tree with height $h = 1$, where the root parameter θ_1 is the prior mean of the actions. Because θ_1 is certain, $w_1 = 0$ and $\sum_{i \in \mathcal{V}} c^{h_i} w_i = \sum_{i=2}^{K+1} w_i = O(K)$.

5.3. Lower Regret Due to Hierarchy

Now we give examples of how the hierarchy can help with reducing regret. To simplify the discussion, we ignore logarithmic factors in the definitions of w_i in Theorem 4. We assume that $\mathcal{T}$ is a balanced b-ary tree with height h; with $K = b^h$ action nodes and $b^{h-\ell}$ nodes at height ℓ. Our discussion is under the assumption that $c = 2$, as derived in Lemma 3. More gains are possible when $c < 2$.

We compare the regret of `HierTS` to classical Thompson sampling (`TS`), which maintains an independent posterior of $\theta_{*,a}$ for each action $a \in \mathcal{A}$. To have a fair comparison, we set the marginal prior variances of all actions in `TS` as in `HierTS`. Specifically, let ψ_a be the path in $\mathcal{T}$ from action node a to the root. Then the marginal prior of action a is $\mathcal{N}(\mu_1, \bar{\sigma}^2_{0,a})$, where $\bar{\sigma}^2_{0,a} = \sum_{i \in \psi_a} \sigma^2_{0,i}$ and μ_1 denotes the hyper-prior mean in (6). The regret of this algorithm can be bounded using Theorem 4, with the only difference that the complexity term becomes $\mathcal{G}_{\texttt{TS}}(n) \approx \sum_{a \in \mathcal{A}} \bar{\sigma}^2_{0,a}$.

Problem 1. We start with a problem where all prior variances are identical, $\sigma^2_{0,i} = 1$ for any $i \in \mathcal{V}$. In this case, all prior variances in `TS` are $\bar{\sigma}^2_{0,a} = h + 1$ and its complexity term is $\mathcal{G}_{\texttt{TS}}(n) = (h+1)b^h$. In `HierTS`, we aggregate the nodes by height and get

$$\mathcal{G}(n) = \sum_{\ell=0}^{h} b^{h-\ell} c^{\ell} = b^h \sum_{\ell=0}^{h} (2/b)^{\ell} \leq \frac{1}{1 - 2/b} b^h \,.$$

Thus `HierTS` improves $\mathcal{G}(n)$ by $\Omega(h)$ when $b > 2$. Since $h = \log_b b^h = \log_b K$, we get $\mathcal{G}_{\texttt{TS}}(n)/\mathcal{G}(n) \approx \log_b K$, and `HierTS` reduces the Bayes regret by a multiplicative factor $\sqrt{\log_b K}$. This argument can be adjusted for $b = 2$ to get a comparable regret to `TS`.

Problem 2. Now we consider a problem where the conditional prior variances in $\mathcal{T}$ double with height, $\sigma^2_{0,i} = 2^{h_i}$, where h_i is the height of node i. This setting is motivated in Section 1. We expect higher statistical gains because the uncertainty of highly-uncertain nodes at higher levels of $\mathcal{T}$ is reduced jointly by all actions. In this problem, all prior variances in TS are $\bar{\sigma}_{0,a}^2 = 2^{h+1}$ and its complexity term is $\mathcal{G}_{\texttt{TS}}(n) = 2^{h+1}b^h$. In comparison, HierTS yields

$$\mathcal{G}(n) = \sum_{\ell=0}^{h} b^{h-\ell}c^\ell 2^\ell \leq \sum_{\ell=0}^{h} b^{h-\ell}4^\ell \leq \frac{1}{1-4/b}b^h\,,$$

where the last step is analogous to Problem 1. So HierTS improves $\mathcal{G}(n)$ by $\Omega(2^{h+1})$ if $b > 4$. Since

$$2^{h+1} = 2 \cdot 2^{\log_b b^h} = 2 \cdot 2^{\log_2 b^h / \log_2 b} = 2K^{\frac{1}{\log_2 b}}\,,$$

we get $\mathcal{G}_{\texttt{TS}}(n)/\mathcal{G}(n) \approx K^{\frac{1}{\log_2 b}}$, and HierTS reduces the Bayes regret by a multiplicative factor $\sqrt{K^{\frac{1}{\log_2 b}}}$. For $b = 5$, this factor would be close to $K^{\frac{1}{4}}$. Therefore, the regret is reduced by a polynomial factor in K.

5.4. Contextual Regret Bound

This section presents a contextual generalization of the regret bound in Theorem 4. We make two assumptions in its derivation. First, the length of the context vector is bounded, and we assume that $\|X_t\|_2 \leq 1$ without loss of generality. Second, $\Sigma_{0,i} = \sigma_{0,i}^2 I_d$. This latter assumption allows us to utilize $\lambda_1(\Sigma_{0,i}) = \lambda_d(\Sigma_{0,i}) = \sigma_{0,i}^2$, which is required by our matrix generalizations of Lemmas 2 and 3. Our regret bound is stated below.

Theorem 5. *For any $\delta > 0$, the Bayes regret of* HierTS *is bounded as*

$$\mathcal{BR}(n) \leq \sqrt{2dn\mathcal{G}(n)\log(1/\delta)} + \sqrt{2/\pi}\sigma_{\max}Kn\delta\,,$$

where $\mathcal{G}(n) = \sum_{i \in \mathcal{V}} c^{h_i} w_i$ and c is a constant defined in Lemma 3. For an action node i, $h_i = 0$ and

$$w_i = \frac{\sigma_{0,i}^2}{\log\left(1 + \frac{\sigma_{0,i}^2}{\sigma^2}\right)} \log\left(1 + \frac{\sigma_{0,i}^2 n}{\sigma^2 d}\right).$$

For an internal node i, $h_i > 0$ and

$$w_i = \frac{\sigma_{0,i}^2}{\log\left(1 + \frac{\sigma_{0,i}^2}{\sigma^2}\right)} \log\left(1 + \sigma_{0,i}^2 \sum_{j \in \mathsf{ch}(i)} \sigma_{0,j}^{-2}\right).$$

The constants are similar to Theorem 4. The main difference is the extra factor of $\sqrt{d}$, because each node is now associated with a learned d-dimensional parameter.

The key insight in the proof is that the context in round t, X_t, is known and fixed. Because all conditional posteriors are Gaussian, posterior sampling in contextual HierTS can be implemented using a tree with scalar nodes, where the conditional posterior of node i is $\mathcal{N}(X_t^\top \hat{\theta}_{t,i}, X_t^\top \hat{\Sigma}_{t,i} X_t)$. As a result, the non-contextual analysis in Theorem 4 can be easily generalized. Lemma 1 changes in that the history H_t includes context X_t. Moreover, the marginal posterior variance of the mean reward of action A_t, $\bar{\sigma}_{t,A_t}^2$, turns into that in the direction of context X_t.

The main technical challenges in the proof are generalizing the variance decomposition in Lemma 3 to covariances (Lemma 9), and extending the posterior update lower bound in Lemma 3 to matrices (Lemma 9). The rest of the proof follows the same outline as that of Theorem 4.

6. Experiments

We compare HierTS to three baselines that either neglect or partially use the tree $\mathcal{T}$. The first baseline is Thompson sampling (TS), which treats each action independently and is introduced in Section 5.3. The second baseline only uses a 2-level hierarchy, namely the root and action nodes, and we call it FlatTS. Its hyper-prior for the root $P_{0,1}$ is the same as in HierTS. For any action $a \in \mathcal{A}$, the conditional prior is $P_{0,a}(\cdot \mid \theta_{*,1}) = \mathcal{N}(\cdot; \theta_{*,1}, \bar{\sigma}_{0,a}^2 - \sigma_{0,1}^2)$, where $\bar{\sigma}_{0,a}^2$ is the marginal prior variance in TS. This baseline mimics existing algorithms for 2-level Gaussian hierarchies with a common root (Kveton et al., 2021; Basu et al., 2021; Hong et al., 2022), and is similar to structured bandits where the actions share a latent parameter (Gupta et al., 2018).

The final baseline is the zooming algorithm of Kleinberg et al. (2008; 2013), which we call Zooming. This is a UCB algorithm where actions are embedded in a metric space, and can be viewed as a standard approach to handling large structured action spaces in bandits (Bubeck et al., 2008; Kleinberg et al., 2013). At a high level, Zooming maintains a set of "active" actions, which are sufficiently apart from each other given the history, and only explores these. The key step in implementing Zooming is constructing a metric space where actions with similar mean rewards are close. We design it as follows. We compute the graph Laplacian of $\mathcal{T}$, represented as an undirected graph, and extract its d eigenvectors $(v_i)_{i=1}^d$ with the smallest eigenvalues. Let $v_{i,j}$ be the j-th entry of v_i. Then, for any action node $a \in \mathcal{A}$, its embedding in the metric space is $(v_{1,a}, \dots, v_{d,a})$. This is a standard approach to deriving a metric space in spectral clustering (Yan et al., 2009).

6.1. Synthetic Experiments

Our first experiments are on a synthetic Gaussian bandit, where we validate theoretical findings from Section 5.3. We experiment with both problems in Section 5.3, which are b-ary trees with height h and $K = b^h$ actions. In Problem 1, the prior variances are constant. In Problem 2, the prior

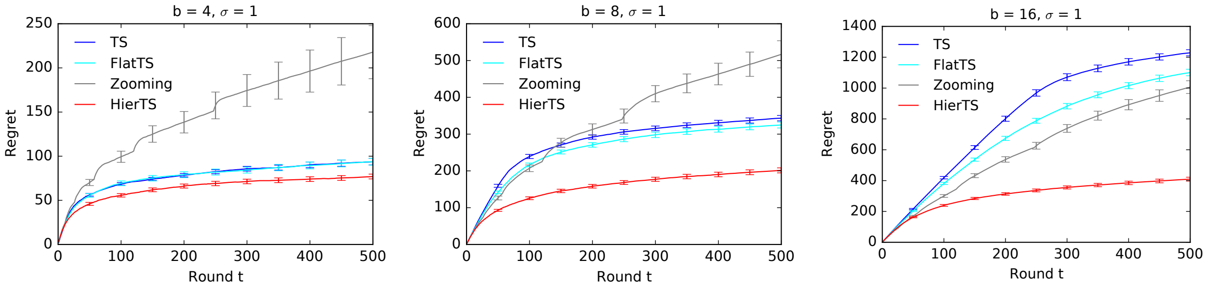

Figure 2. Regret of `HierTS` on synthetic bandit problems with varying branching factor b.

variances double with height. In both problems, the mean of the hyper-prior is $\mu_1 = 0$, and the reward of action a is $\theta_{*,a}$ with variance $\sigma^2 = 1$.

We start with Problem 2, where the tree height is $h = 2$ and we vary the branching factor b. All algorithms are run for $n = 500$ rounds and evaluated by the Bayes regret on 100 independent samples of Θ_*. Its mean estimate and standard error are reported in Figure 2. For all values of b, `HierTS` outperforms all baselines. `Zooming` is not competitive with the other baselines when the branching factor b is low. The poor performance of `Zooming` is because of its generality. Specifically, since it can be applied to any smooth reward function, its regret in d dimensions scales as $\tilde{O}(n^{1-\frac{1}{2+d}})$. In Figure 2, we only show the results for `Zooming` where $d = K$. We experimented with $d < K$ and observed linear regret, because some action nodes with a common parent had identical features.

In the next experiment, we consider both Problems 1 and 2. The branching factor is fixed at $b = 2$ and we vary the tree height h. All algorithms are run for $n = 500$ rounds on 100 independent samples of Θ_*. The reduction in the Bayes regret of `HierTS` and `FlatTS` is measured as the ratio of the TS regret over the regret in question. In Figure 3a, we plot the ratios for Problem 1. Section 5.3 suggests that the regret decreases by $\Omega(\sqrt{h})$. Our plot confirms this. In Figure 3b, we plot the ratios for Problem 2. Section 5.3 suggests that the regret decreases by $\Omega(2^{h/2})$. We confirm this trend.

6.2. Multi-Label Image Classification

The last experiment is on a multi-label image classification problem with linear rewards. All compared algorithms are implemented analogously to Section 6.1, except for variances being replaced with covariances. We use the CIFAR-100 dataset (Krizhevsky, 2009), with 60 000 images of size 32×32. There are 50 000 training and 10 000 test images. Each image belongs to one of 100 classes (labels) and 20 super-classes, each consisting of 5 classes. Each image is represented by a $d = 10$ dimensional feature vector, which we obtain by downsampling a 100-dimensional feature vector. That feature vector is an embedding computed by an EfficientNet-L2 network applied to the image (Tan & Le, 2019; Xie et al., 2020; Foret et al., 2021). The network is a convolutional neural network pretrained on both ImageNet (Russakovsky et al., 2015) and unlabeled JFT-300M (Sun et al., 2017), and fine-tuned on the CIFAR-100 training set.

We randomly select 5 super-classes and their corresponding $K = 25$ classes become actions. The test and training sets are restricted to these classes. Our bandit problem is defined as follows. For each action a, $\theta_{*,a}$ is the mean feature vector of test images in class a. In round t, the context X_t is the feature vector of a random image from the test set and the reward of action A_t is $Y_t \sim \mathcal{N}(X_t^\top \theta_{*,A_t}, 0.5^2)$. Therefore, the mean reward is maximized whenever the true class is chosen. Finally, we build a 3-level hierarchy $\mathcal{T}$ as follows. The hyper-prior of the root is a multivariate Gaussian fitted to all *training images*, $P_{0,1} = \mathcal{N}(\mu_1, \Sigma_{0,1})$. The nodes at height 1 are the 5 super-classes and their conditional priors are $P_{0,i}(\cdot \mid \theta_{*,1}) = \mathcal{N}(\cdot; \theta_{*,1}, \Sigma_{0,i})$, where $\Sigma_{0,i}$ is fitted to the training images of super-class i. Finally, the nodes at height 0 correspond to actions and their conditional priors are $P_{0,a}(\cdot \mid \theta_{*,\mathrm{pa}(a)}) = \mathcal{N}(\cdot; \theta_{*,\mathrm{pa}(a)}, \Sigma_{0,a})$, where $\Sigma_{0,a}$ is fitted to the training images of class a.

We report the mean and standard error of the regret over 10 runs in Figure 4. We observe again that `HierTS` performs better than all baselines. We do not consider `Zooming` due to its poor performance earlier (Figure 2). Note that the true model parameters of $\mathcal{T}$, namely μ_1 and $\Sigma_{0,i}$, are unknown; and we estimate them from training images. Therefore, this experiment shows that even when we relax the assumption that they are known, it is beneficial to estimate them, and use the structure of $\mathcal{T}$.

7. Related Work

Thompson sampling algorithms have been widely applied to contextual bandits because of their computational efficiency and strong empirical performance (Chu et al., 2011; Chapelle & Li, 2012; Abbasi-Yadkori et al., 2011). Russo & Van Roy (2014) proved first Bayes regret bounds for TS. Our proposed algorithm `HierTS` extends TS to tree hierarchies. TS with a 2-level hierarchy over tasks was applied

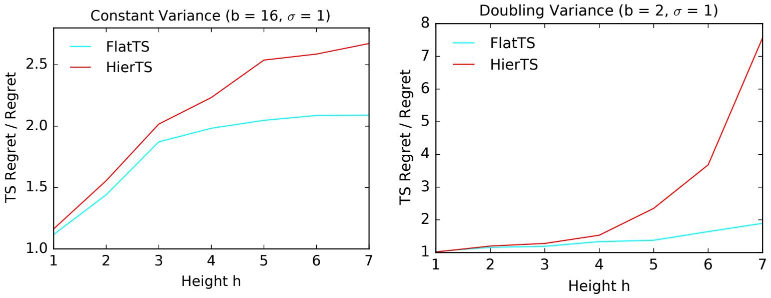

Figure 3. Improvement in the `HierTS` regret when the prior variance (a) is constant or (b) increases with height, as a function of the tree height h.

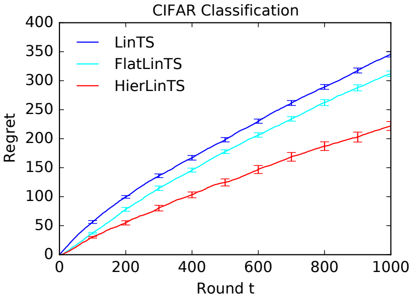

Figure 4. Regret of `HierTS` on a CIFAR-100 image classification problem.

and analyzed in both meta-learning and multi-task learning (Kveton et al., 2021; Basu et al., 2021; Wan et al., 2021; Hong et al., 2022). The main difference in our work is that we move from 2-level hierarchies to an arbitrary depth, and develop both algorithmic and theory foundations for this setting. Our analysis extends the variance decompositions proposed in Hong et al. (2022) to trees. Alternatively, information theory could be used to derive Bayes regret bounds (Russo & Van Roy, 2016; Lu & Van Roy, 2019; Basu et al., 2021), but we are unaware of any for trees.

Our problem has a structured action space. One structure studied by prior works is a shared latent parameter among all actions (Tirinzoni et al., 2020; Lattimore & Munos, 2014; Gupta et al., 2018). This can be viewed as an instance of our setting with a 2-level hierarchy. In latent bandits, the parameter is a discrete variable (Maillard & Mannor, 2014; Hong et al., 2020). Recent works also applied approximate TS to more complex structures (Gopalan et al., 2014; Yu et al., 2020b). Such algorithms are general, but can only be analyzed in limited settings under strong assumptions. We consider a special tree structure, where we can derive and analyze an exact algorithm. Another recent work that applies trees to bandits is Majzoubi et al. (2020), who proposed a general reduction-based algorithm for contextual bandits with a continuous action space, where the policies are represented by a tree.

The most related work is Sen et al. (2021), who also studied contextual bandits with a tree hierarchy over actions. Both of our works rely on a hierarchy of regressors, motivated by multi-label classification (Prabhu et al., 2018; Yu et al., 2020a). However, the works differ in several key aspects. First, we study a stochastic variant of the problem, where the tree nodes are associated with prior distributions rather than fixed centers and radii, as in Sen et al. (2021). In fact, Sen et al. (2021) did not model the statistical uncertainty at all. Second, we propose a TS algorithm using novel recursive derivations of the posterior. Sen et al. (2021) proposed a greedy strategy and beam search to avoid evaluation of all actions. In Section 3, we discuss how similar improvements can be incorporated in `HierTS`. Finally, our Bayesian analysis reveals structural properties that imply low regret. The low regret in Sen et al. (2021) is attained by making an assumption on the regression oracle, and can grow linearly when the oracle is imperfect.

8. Conclusions

In many practical problems, the action space is large and a good generalization over actions is not obvious. Motivated by this, we study a contextual bandit problem with a *deep hierarchy* over actions. To solve the problem, we propose hierarchical Thompson sampling (`HierTS`), which can be implemented exactly and efficiently in Gaussian models. We prove Bayes regret bounds for `HierTS` that quantify its increased statistical efficiency over vanilla TS and validate this experimentally. We also apply `HierTS` to a challenging classification problem with label hierarchy.

Our work is a major step towards bandit algorithms with rich graphical models. Its limitations can be addressed by future works. For instance, a frequentist regret analysis is possible and would only differ in Lemma 1. The rest of the proof, which captures the structure of our problem, is a worst-case argument. Second, we believe that our method can be generalized beyond Gaussian trees. As discussed in Section 3.2, exact posterior sampling is challenging under the constraint of computational efficiency; but many tractable approximations exist. When exact sampling is possible, we believe that our proofs can be extended to general exponential-family distributions. Another direction for future work is an extension to directed acyclic graphs (DAGs). The nodes in DAGs can be ordered, and therefore similar recursions to Sections 4 and 5 can be established.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

Agrawal, S. and Goyal, N. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceeding of the 25th Annual Conference on Learning Theory*, pp. 39.1–39.26, 2012.

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Basu, S., Kveton, B., Zaheer, M., and Szepesvari, C. No regrets for learning the prior in bandits. In *Advances in Neural Information Processing Systems 34*, 2021.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvari, C. Online optimization in X-armed bandits. In *Advances in Neural Information Processing Systems 21*, pp. 201–208, 2008.

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pp. 2249–2257, 2012.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Combes, R., Magureanu, S., Proutiere, A., and Laroche, C. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 2015.

Doucet, A., de Freitas, N., and Gordon, N. *Sequential Monte Carlo Methods in Practice*. Springer, New York, NY, 2001.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Gopalan, A., Mannor, S., and Mansour, Y. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 100–108, 2014.

Gupta, S., Chaudhari, S., Mukherjee, S., Joshi, G., and Yagan, O. A unified approach to translate classical bandit algorithms to the structured bandit setting. *CoRR*, abs/1810.08164, 2018. URL https://arxiv.org/abs/1810.08164.

Hong, J., Kveton, B., Zaheer, M., Chow, Y., Ahmed, A., and Boutilier, C. Latent bandits revisited. In *Advances in Neural Information Processing Systems 33*, 2020.

Hong, J., Kveton, B., Zaheer, M., and Ghavamzadeh, M. Hierarchical Bayesian bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Kawale, J., Bui, H., Kveton, B., Tran-Thanh, L., and Chawla, S. Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems 28*, pp. 1297–1305, 2015.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008.

Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *CoRR*, abs/1312.1277, 2013. URL https://arxiv.org/abs/1312.1277.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-W., Mladenov, M., Boutilier, C., and Szepesvari, C. Meta-Thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Lattimore, T. and Munos, R. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems 27*, pp. 550–558, 2014.

Li, L., Chu, W., Langford, J., and Schapire, R. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

Li, S., Wang, B., Zhang, S., and Chen, W. Contextual combinatorial cascading bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1245–1253, 2016.

Lindley, D. and Smith, A. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18, 1972.

Lu, X. and Van Roy, B. Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, 2019.

Maillard, O.-A. and Mannor, S. Latent bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 136–144, 2014.

Majzoubi, M., Zhang, C., Chari, R., Krishnamurthy, A., Langford, J., and Slivkins, A. Efficient contextual bandits with continuous actions. In *Advances in Neural Information Processing Systems 33*, 2020.

Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., and Varma, M. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*, 2018.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Russo, D. and Van Roy, B. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

Sen, R., Shanmugam, K., Kocaoglu, M., Dimakis, A., and Shakkottai, S. Contextual bandits with latent confounders: An NMF approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Sen, R., Rakhlin, A., Ying, L., Kidambi, R., Foster, D., Hill, D., and Dhillon, I. Top-k extreme contextual bandits with arm hierarchy. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision*, 2017.

Tan, M. and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, 2019.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Tirinzoni, A., Lazaric, A., and Restelli, M. A novel confidence-based algorithm for structured bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

Wan, R., Ge, L., and Song, R. Metadata-based multi-task bandits with Bayesian hierarchical models. In *Advances in Neural Information Processing Systems 34*, 2021.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. Self-training with noisy student improves ImageNet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yan, D., Huang, L., and Jordan, M. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

Yu, H.-F., Zhong, K., Zhang, J., Chang, W.-C., and Dhillon, I. PECOS: Prediction for enormous and correlated output spaces. *CoRR*, abs/2010.05878, 2020a. URL http://arxiv.org/abs/2010.05878.

Yu, T., Kveton, B., Wen, Z., Zhang, R., and Mengshoel, O. Graphical models meet bandits: A variational Thompson sampling approach. In *Proceedings of the 37th International Conference on Machine Learning*, 2020b.

Yue, Y. and Guestrin, C. Linear submodular bandits and their application to diversified retrieval. In *Advances in Neural Information Processing Systems 24*, pp. 2483–2491, 2011.

Zhang, Y. and Yang, Q. A survey on multi-task learning. *CoRR*, abs/1707.08114, 2017. URL http://arxiv.org/abs/1707.08114.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. Cascading bandits for large-scale recommendation problems. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 2016.

A. Posterior Derivations

This section contains our posterior derivations. We adopt the convention that $\Lambda = \Sigma^{-1}$, where Λ is the precision matrix for covariance Σ.

A.1. Multi-Armed Bandit Posterior

The proof is by induction. We start with the inductive step.

Lemma 6. *Fix an internal node j. Let $\mathsf{pa}(j) = i$ and $C = \mathsf{ch}(j)$. Let*

$$\mathbb{P}\left(H_{t,j}, \theta_{*,j} = \theta \,|\, \theta_{*,i} = \theta_i\right) \propto \exp\left[-\frac{1}{2}\left(\sigma_0^{-2}(\theta - \theta_i)^2 + \sum_{k \in C} \sigma_k^{-2}(\theta - \theta_k)^2\right)\right], \tag{13}$$

where θ_k, σ_k are the parameters of $k \in C$. Then

$$\mathbb{P}\left(H_{t,j} \,|\, \theta_{*,i} = \theta_i\right) \propto \exp\left[-\frac{1}{2}\tilde{\sigma}_j^{-2}(\theta_i - \tilde{\theta}_j)^2\right]$$

for $\tilde{\sigma}_j^2 = \sigma_0^2 + (\sum_{k \in C} \sigma_k^{-2})^{-1}$ and $\tilde{\theta}_j = (\sum_{k \in C} \sigma_k^{-2})^{-1} \sum_{k \in C} \sigma_k^{-2} \theta_k$.

Proof. Let $s = \sigma_0^{-2} + \sum_{k \in C} \sigma_k^{-2}$ and $v = \sigma_0^{-2}\theta_i + \sum_{k \in C} \sigma_k^{-2}\theta_k$. We start with completing the square of θ,

$$\begin{aligned}
\log \mathbb{P}\left(H_{t,j}, \theta_{*,j} = \theta \,|\, \theta_{*,i} = \theta_i\right) &\propto \sigma_0^{-2}(\theta - \theta_i)^2 + \sum_{k \in C} \sigma_k^{-2}(\theta - \theta_k)^2 \\
&\propto s\theta^2 - 2\theta\left(\sigma_0^{-2}\theta_i + \sum_{k \in C} \sigma_k^{-2}\theta_k\right) + \sigma_0^{-2}\theta_i^2 \\
&= s(\theta^2 - 2\theta s^{-1} v + s^{-2} v^2) + \sigma_0^{-2}\theta_i^2 - s^{-1} v^2 \\
&= s(\theta - s^{-1} v)^2 + \sigma_0^{-2}\theta_i^2 - s^{-1} v^2 \,.
\end{aligned}$$

In the second step, we omit constants in θ and θ_i. Since we got a quadratic form in θ, we know that

$$\int_\theta \mathbb{P}\left(H_{t,j}, \theta_{*,j} = \theta \,|\, \theta_{*,i} = \theta_i\right) \mathrm{d}\theta \propto \exp\left[-\frac{1}{2}(\sigma_0^{-2}\theta_i^2 - s^{-1} v^2)\right].$$

Let $\hat{s} = \sigma_0^{-2} - \sigma_0^{-4} s^{-1}$. Now we complete the square of θ_i,

$$\begin{aligned}
\sigma_0^{-2}\theta_i^2 - s^{-1} v^2 = \sigma_0^{-2}\theta_i^2 - s^{-1}\left(\sigma_0^{-2}\theta_i + \sum_{k \in C} \sigma_k^{-2}\theta_k\right)^2 &= \sigma_0^{-2}\theta_i^2 - \sigma_0^{-4} s^{-1}\left(\theta_i + \sigma_0^2 \sum_{k \in C} \sigma_k^{-2}\theta_k\right)^2 \\
\propto \hat{s}\left(\theta_i^2 - 2\theta_i \hat{s}^{-1} \sigma_0^{-2} s^{-1} \sum_{k \in C} \sigma_k^{-2}\theta_k\right) &\propto \hat{s}\left(\theta_i - \hat{s}^{-1}\sigma_0^{-2} s^{-1} \sum_{k \in C} \sigma_k^{-2}\theta_k\right)^2 .
\end{aligned}$$

In the last two steps, we omit constants in θ_i. Finally, note that

$$\begin{aligned}
\hat{s} &= \frac{\sigma_0^{-2}(s - \sigma_0^{-2})}{s} = (\sigma_0^2 + (s - \sigma_0^{-2})^{-1})^{-1} \,, \\
\hat{s}^{-1}\sigma_0^{-2} s^{-1} &= \frac{s}{\sigma_0^{-2}(s - \sigma_0^{-2})}\sigma_0^{-2} s^{-1} = (s - \sigma_0^{-2})^{-1} \,.
\end{aligned}$$

This completes the proof, for $\tilde{\theta}_j = (s - \sigma_0^{-2})^{-1} \sum_{k \in C} \sigma_k^{-2}\theta_k$ and $\tilde{\sigma}_j^2 = \sigma_0^2 + (s - \sigma_0^{-2})^{-1}$. □

At an action node j, (13) holds for $\sigma_k = \sigma$ and $\theta_k = Y_k$, where Y_k is an observation k of node j and σ is observation noise. This is the base case of our proof by induction.

A.2. Linear Bandit Posterior

The proof is by induction. We start with the inductive step.

Lemma 7. *Fix an internal node j. Let $\mathsf{pa}(j) = i$ and $C = \mathsf{ch}(j)$. Let*

$$\mathbb{P}\left(H_{t,j}, \theta_{*,j} = \theta \,|\, \theta_{*,i} = \theta_i\right) \propto \exp\left[-\frac{1}{2}\left((\theta - \theta_i)^\top \Lambda_0 (\theta - \theta_i) + \sum_{k \in C} (\theta - \theta_k)^\top \Lambda_k (\theta - \theta_k)\right)\right], \tag{14}$$

where θ_k, Λ_k are the parameters of $k \in C$. Then

$$\mathbb{P}\left(H_{t,j} \,|\, \theta_{*,i} = \theta_i\right) \propto \exp\left[-\frac{1}{2}(\theta_i - \tilde{\theta}_j)^\top \tilde{\Lambda}_j (\theta_i - \tilde{\theta}_j)\right]$$

for $\tilde{\Lambda}_j^{-1} = \Lambda_0^{-1} + (\sum_{k \in C} \Lambda_k)^{-1}$ and $\tilde{\theta}_j = (\sum_{k \in C} \Lambda_k)^{-1} \sum_{k \in C} \Lambda_k \mu_k$.

Proof. Let $S = \Lambda_0 + \sum_{k \in C} \Lambda_k$ and $V = \Lambda_0 \theta_i + \sum_{k \in C} \Lambda_k \theta_k$. We start with completing the square of θ,

$$\begin{aligned}
\log \mathbb{P}\left(H_{t,j}, \theta_{*,j} = \theta \,|\, \theta_{*,i} = \theta_i\right) &\propto (\theta - \theta_i)^\top \Lambda_0 (\theta - \theta_i) + \sum_{k \in C} (\theta - \theta_k)^\top \Lambda_k (\theta - \theta_k) \\
&\propto \theta^\top S \theta - 2\theta^\top \left(\Lambda_0 \theta_i + \sum_{k \in C} \Lambda_k \theta_k\right) + \theta_i^\top \Lambda_0 \theta_i \\
&= \theta^\top S (\theta - 2 S^{-1} V) + \theta_i^\top \Lambda_0 \theta_i \\
&= (\theta - S^{-1} V)^\top S (\theta - S^{-1} V) + \theta_i^\top \Lambda_0 \theta_i - V^\top S^{-1} V \,.
\end{aligned}$$

In the second step, we omit constants in θ and θ_i. Since we got a quadratic form in θ, we know that

$$\int_\theta \mathbb{P}\left(H_{t,j}, \theta_{*,j} = \theta \,|\, \theta_{*,i} = \theta_i\right) \mathrm{d}\theta \propto \exp\left[-\frac{1}{2}(\theta_i^\top \Lambda_0 \theta_i - V^\top S^{-1} V)\right] .$$

Let $\hat{S} = \Lambda_0 - \Lambda_0 S^{-1} \Lambda_0$. Now we complete the square of θ_i,

$$\begin{aligned}
\theta_i^\top \Lambda_0 \theta_i - V^\top S^{-1} V &= \theta_i^\top \Lambda_0 \theta_i - \left(\Lambda_0 \theta_i + \sum_{k \in C} \Lambda_k \theta_k\right)^\top S^{-1} \left(\Lambda_0 \theta_i + \sum_{k \in C} \Lambda_k \theta_k\right) \\
&\propto \theta_i^\top \hat{S} \left(\theta_i - 2 \hat{S}^{-1} \Lambda_0 S^{-1} \sum_{k \in C} \Lambda_k \theta_k\right) \\
&\propto \left(\theta_i - \hat{S}^{-1} \Lambda_0 S^{-1} \sum_{k \in C} \Lambda_k \theta_k\right)^\top \hat{S} \left(\theta_i - \hat{S}^{-1} \Lambda_0 S^{-1} \sum_{k \in C} \Lambda_k \theta_k\right) .
\end{aligned}$$

In the last two steps, we omit constants in θ_i. Finally, by the Woodbury matrix identity, we have

$$\begin{aligned}
\hat{S} = \Lambda_0 - \Lambda_0 S^{-1} \Lambda_0 &= (\Lambda_0^{-1} + (S - \Lambda_0)^{-1})^{-1} \,, \\
\hat{S}^{-1} \Lambda_0 S^{-1} = (\Lambda_0 - \Lambda_0 S^{-1} \Lambda_0)^{-1} \Lambda_0 S^{-1} &= (S - \Lambda_0)^{-1} \,.
\end{aligned}$$

This completes the proof, for $\tilde{\theta}_j = (S - \Lambda_0)^{-1} \sum_{k \in C} \Lambda_k \mu_k$ and $\tilde{\Lambda}_j = (\Lambda_0^{-1} + (S - \Lambda_0)^{-1})^{-1}$. □

For any node j, note that (14) can be written as

$$\mathbb{P}\left(H_{t,j}, \theta_{*,j} = \theta \,|\, \theta_{*,i} = \theta_i\right) \propto \exp\left[-\frac{1}{2}\left((\theta - \theta_i)^\top \Lambda_0 (\theta - \theta_i) + \sum_{k \in C} \theta^\top \Lambda_k \theta - 2 \sum_{k \in C} \theta^\top \Lambda_k \theta_k)\right)\right] ,$$

when constants in θ and θ_i are omitted. Then, at an action node, $\Lambda_k = \sigma^{-2} X_k^\top X_k$ and $\Lambda_k \theta_k = \sigma^{-2} X_k Y_k$, where Y_k is an observation k of node j at feature vector X_k and σ is observation noise. This is the base case of our proof by induction.

B. Regret Bound Proofs

This section contains proofs of our regret bound and supporting lemmas.

B.1. Proof of Lemma 1

Fix round t. Let $\mathbb{P}\left(\Theta \mid H_t\right) = \mathcal{N}(\Theta; \bar{\Theta}_t, \bar{\Sigma}_t)$ be the joint posterior distribution of all action node parameters $\Theta \in \mathbb{R}^K$, with mean $\bar{\Theta}_t \in \mathbb{R}^K$ and covariance $\bar{\Sigma}_t \in \mathbb{R}^{K \times K}$. Let $A_t \in \{0, 1\}^K$ and $A_* \in \{0, 1\}^K$ be indicator vectors of the taken action in round t and the optimal action, respectively. Each action is associated with one leaf node.

Since $\bar{\Theta}_t$ is deterministic given H_t, and A_* and A_t are i.i.d. given H_t, we have

$$\mathbb{E}\left[A_*^\top \Theta_* - A_t^\top \Theta_*\right] = \mathbb{E}\left[\mathbb{E}\left[A_*^\top (\Theta_* - \bar{\Theta}_t) \,\middle|\, H_t\right]\right] + \mathbb{E}\left[\mathbb{E}\left[A_t^\top (\bar{\Theta}_t - \Theta_*) \,\middle|\, H_t\right]\right] .$$

Moreover, $\Theta_* - \bar{\Theta}_t$ is a zero-mean random vector independent of A_t, and thus $\mathbb{E}\left[A_t^\top (\bar{\Theta}_t - \Theta_*) \,\middle|\, H_t\right] = 0$. So we only need to bound the first term above. Let

$$E_t = \left\{\forall a \in \mathcal{A} : |a^\top (\Theta_* - \bar{\Theta}_t)| \leq \sqrt{2 \log(1 / \delta)} \|a\|_{\bar{\Sigma}_t}\right\}$$

be the event that all high-probability confidence intervals hold. Fix history H_t. Then by the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[A_*^\top (\Theta_* - \bar{\Theta}_t) \,\middle|\, H_t\right] \leq \sqrt{2 \log(1 / \delta)}\, \mathbb{E}\left[\|A_*\|_{\bar{\Sigma}_t} \,\middle|\, H_t\right] + \mathbb{E}\left[A_*^\top (\Theta_* - \bar{\Theta}_t) \mathbb{1}\{\bar{E}_t\} \,\middle|\, H_t\right] .$$

Now note that for any action a, $a^\top (\Theta_* - \bar{\Theta}_t) / \|a\|_{\bar{\Sigma}_t}$ is a standard normal variable. It follows that

$$\mathbb{E}\left[A_*^\top (\Theta_* - \bar{\Theta}_t) \mathbb{1}\{\bar{E}_t\} \,\middle|\, H_t\right] \leq 2 \sum_{a \in \mathcal{A}} \|a\|_{\bar{\Sigma}_t} \frac{1}{\sqrt{2 \pi}} \int_{u = \sqrt{2 \log(1 / \delta)}}^\infty u \exp\left[- \frac{u^2}{2}\right] \mathrm{d}u \leq \sqrt{\frac{2}{\pi}} \sigma_{\max} K \delta \,,$$

where we use that $\bar{E}_t$ implies $\|\bar{\Sigma}_t^{- \frac{1}{2}} (\Theta_* - \bar{\Theta}_t)\|_\infty \geq \sqrt{2 \log(1 / \delta)}$. Now we combine all inequalities and have

$$\mathbb{E}\left[A_*^\top (\Theta_* - \bar{\Theta}_t) \,\middle|\, H_t\right] \leq \sqrt{2 \log(1 / \delta)}\, \mathbb{E}\left[\|A_t\|_{\bar{\Sigma}_t} \,\middle|\, H_t\right] + \sqrt{\frac{2}{\pi}} \sigma_{\max} K \delta \,.$$

We also used that A_t and A_* are i.i.d. given H_t.

Since the above bound holds for any history H_t, we combine everything and get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t = 1}^n A_*^\top \Theta_* - A_t^\top \Theta_*\right]
& \leq \sqrt{2 \log(1 / \delta)}\, \mathbb{E}\left[\sum_{t = 1}^n \|A_t\|_{\bar{\Sigma}_t}\right] + \sqrt{\frac{2}{\pi}} \sigma_{\max} K n \delta \\
& \leq \sqrt{2 n \log(1 / \delta)}\, \mathbb{E}\left[\sqrt{\sum_{t = 1}^n \|A_t\|_{\bar{\Sigma}_t}^2}\right] + \sqrt{\frac{2}{\pi}} \sigma_{\max} K n \delta \\
& \leq \sqrt{2 n \log(1 / \delta)} \sqrt{\mathbb{E}\left[\sum_{t = 1}^n \|A_t\|_{\bar{\Sigma}_t}^2\right]} + \sqrt{\frac{2}{\pi}} \sigma_{\max} K n \delta \,.
\end{aligned}$$

The second step uses the Cauchy-Schwarz inequality and the third step uses the concavity of the square root.

Since a is an indicator vector, $\|A_t\|_{\bar{\Sigma}_t}^2 = \bar{\sigma}_{t, A_t}^2$ is the marginal posterior variance of the mean reward of action A_t in round t. Likewise, $\sigma_{\max}$ is the maximum marginal prior width of the mean reward at an action node. This concludes the proof.

B.2. Proof of Lemma 2

Since round t is fixed, we write L instead of L_t and refer to node $\psi_t(i)$ by i for any $i \in [L_t]$.

Fix any node i. By the total variance decomposition, where we introduce a parent parameter $\theta_{t, i - 1}$,

$$\mathrm{var}\left[\theta_{t, i} \mid H_t\right] = \mathbb{E}\left[\hat{\sigma}_{t, i}^2 \,\middle|\, H_t\right] + \mathrm{var}\left[\hat{\theta}_{t, i} \,\middle|\, H_t\right] .$$

For Gaussian random variables, $\hat{\sigma}_{t,i}^2$ is independent of $\theta_{t,i-1}$, as shown in (7). Therefore,

$$\mathbb{E}\left[\hat{\sigma}_{t,i}^2 \,\middle|\, H_t\right] = \hat{\sigma}_{t,i}^2\,.$$

For the second term, as shown in (7), $\hat{\theta}_{t,i} = \hat{\sigma}_{t,i}^2(\sigma_{0,i}^{-2}\theta_{t,i-1} + c)$, where c is a constant conditioned on H_t. Therefore,

$$\mathrm{var}\left[\hat{\theta}_{t,i} \,\middle|\, H_t\right] = \frac{\hat{\sigma}_{t,i}^4}{\sigma_{0,i}^4}\mathrm{var}\left[\theta_{t,i-1} \,|\, H_t\right]\,.$$

Now we chain all identities for node i and get

$$\mathrm{var}\left[\theta_{t,i} \,|\, H_t\right] = \hat{\sigma}_{t,i}^2 + \frac{\hat{\sigma}_{t,i}^4}{\sigma_{0,i}^4}\mathrm{var}\left[\theta_{t,i-1} \,|\, H_t\right]\,.$$

Finally, we apply the above identity recursively, from node L up to the root, and get our claim,

$$\begin{aligned}\mathrm{var}\left[\theta_{t,L} \,|\, H_t\right] &= \hat{\sigma}_{t,L}^2 + \frac{\hat{\sigma}_{t,L}^4}{\sigma_{0,L}^4}\mathrm{var}\left[\theta_{t,L-1} \,|\, H_t\right] = \hat{\sigma}_{t,L}^2 + \frac{\hat{\sigma}_{t,L}^4}{\sigma_{0,L}^4}\hat{\sigma}_{t,L-1}^2 + \frac{\hat{\sigma}_{t,L}^4}{\sigma_{0,L}^4}\frac{\hat{\sigma}_{t,L-1}^4}{\sigma_{0,L-1}^4}\mathrm{var}\left[\theta_{t,L-2} \,|\, H_t\right] \\ &= \sum_{i=1}^{L}\left(\prod_{j=i+1}^{L}\frac{\hat{\sigma}_{t,j}^4}{\sigma_{0,j}^4}\right)\hat{\sigma}_{t,i}^2\,.\end{aligned}$$

This completes the proof.

B.3. Proof of Lemma 3

Since round t is fixed, we write L instead of L_t and refer to node $\psi_t(i)$ by i for any $i \in [L_t]$.

For an action node i, the claim holds trivially since $\hat{\sigma}_{t+1,i}^{-2} - \hat{\sigma}_{t,i}^{-2} = \sigma^{-2}$. The rest of the proof is for internal nodes. We start with proving that

$$\sigma^{-2} \geq \tilde{\sigma}_{t+1,i}^{-2} - \tilde{\sigma}_{t,i}^{-2} \geq \sigma^{-2}\left(\prod_{j=i}^{L}\frac{\hat{\sigma}_{t,i}^4}{\sigma_{0,i}^4}\frac{1}{1+\sigma^{-2}\sigma_{0,i}^2}\right) \tag{15}$$

holds for any node i. The proof is by induction.

The basis of the induction is that the claim holds for action nodes. Let node $i = L$ be an action node, with n observations by round t. We apply (8) and get

$$\begin{aligned}\tilde{\sigma}_{t+1,i}^{-2} - \tilde{\sigma}_{t,i}^{-2} &= (\sigma_{0,i}^2 + \sigma^2/(n+1))^{-1} - (\sigma_{0,i}^2 + \sigma^2/n)^{-1} = \sigma_{0,i}^{-4}[(\sigma_{0,i}^{-2} + \sigma^{-2}n)^{-1} - (\sigma_{0,i}^{-2} + \sigma^{-2}(n+1))^{-1}] \\ &= \sigma_{0,i}^{-4}(\sigma_{0,i}^{-2} + \sigma^{-2}n)^{-2}\frac{\sigma^{-2}}{1+\sigma^{-2}(\sigma_{0,i}^{-2} + \sigma^{-2}n)^{-1}} = \frac{\hat{\sigma}_{t,i}^4}{\sigma_{0,i}^4}\frac{\sigma^{-2}}{1+\sigma^{-2}\hat{\sigma}_{t,i}^2}\,.\end{aligned}$$

The second and third equalities are by the Woodbury and Sherman-Morrison formulas, respectively, applied to scalars. The lower bound follows from $\hat{\sigma}_{t,i}^2 \leq \sigma_{0,i}^2$. The upper bound uses that $\hat{\sigma}_{t,i}^4/\sigma_{0,i}^4 \leq 1$ and $1 + \sigma^{-2}\hat{\sigma}_{t,i}^2 \geq 1$.

In the inductive step, we assume that (15) holds for node $i+1$ and prove it for node i. Note that node $i+1$ is the only child of node i where the posterior between rounds t and $t+1$ changes. Let $s = \sum_{j \in \mathsf{ch}(i)} \tilde{\sigma}_{t,j}^{-2}$ and $\varepsilon = \tilde{\sigma}_{t+1,i+1}^{-2} - \tilde{\sigma}_{t,i+1}^{-2}$. Now we apply (9) and get

$$\begin{aligned}\tilde{\sigma}_{t+1,i}^{-2} - \tilde{\sigma}_{t,i}^{-2} &= (\sigma_{0,i}^2 + (s+\varepsilon)^{-1})^{-1} - (\sigma_{0,i}^2 + s^{-1})^{-1} = \sigma_{0,i}^{-4}[(\sigma_{0,i}^{-2} + s)^{-1} - (\sigma_{0,i}^{-2} + s + \varepsilon)^{-1}] \\ &= \sigma_{0,i}^{-4}(\sigma_{0,i}^{-2} + s)^{-2}\frac{\varepsilon}{1 + (\sigma_{0,i}^{-2} + s)^{-1}\varepsilon} = \frac{\hat{\sigma}_{t,i}^4}{\sigma_{0,i}^4}\frac{\varepsilon}{1+\hat{\sigma}_{t,i}^2\varepsilon}\,.\end{aligned}$$

As in the proof for the action node, the second and third equalities are by the Woodbury and Sherman-Morrison formulas, respectively, applied to scalars. The lower bound follows from $\hat{\sigma}_{t,i}^2 \leq \sigma_{0,i}^2$ and $\varepsilon \leq \sigma^{-2}$, where the latter is by the inductive argument. The upper bound uses that $\hat{\sigma}_{t,i}^4/\sigma_{0,i}^4 \leq 1$, $1 + \hat{\sigma}_{t,i}^2\varepsilon \geq 1$, and $\varepsilon \leq \sigma^{-2}$.

To complete the proof, we have from (7) that

$$\hat{\sigma}_{t+1,i}^{-2} - \hat{\sigma}_{t,i}^{-2} = \tilde{\sigma}_{t+1,i+1}^{-2} - \tilde{\sigma}_{t,i+1}^{-2}$$

holds for any internal node i. Finally, note that $1 + \sigma^{-2}\sigma_{0,i}^2 \leq 1 + \sigma^{-2}\sigma_{0,\max}^2 = c$.

B.4. Proof of Theorem 4

First, we apply Lemma 1 and get that

$$\mathcal{BR}(n) \leq \sqrt{2n\mathcal{V}(n)\log(1/\delta)} + \sqrt{2/\pi}\sigma_{\max}Kn\delta\,,$$

where $\mathcal{V}(n) = \mathbb{E}\left[\sum_{t=1}^n \bar{\sigma}_{t,A_t}^2\right]$ and $\bar{\sigma}_{t,A_t}^2$ is the marginal posterior variance in node A_t in round t. We derive a worst-case upper bound on $\mathcal{V}(n)$ next.

We start with a worst-case upper in any round t. Since round t is fixed, we write L instead of L_t and refer to node $\psi_t(i)$ by i for any $i \in [L_t]$. Let $s_i = \prod_{j=i+1}^{L} \frac{\hat{\sigma}_{t,j}^2}{\sigma_{0,j}^2}$. Then by Lemma 2,

$$\bar{\sigma}_{t,A_t}^2 = \sigma^2 \frac{\bar{\sigma}_{t,A_t}^2}{\sigma^2} = \sigma^2 \sum_{i=1}^{L} s_i^2 \frac{\hat{\sigma}_{t,i}^2}{\sigma^2} \leq \sigma^2 \sum_{i=1}^{L} c_i \log\left(1 + s_i^2 \frac{\hat{\sigma}_{t,i}^2}{\sigma^2}\right)\,,$$

where c_i is an upper bound at node i defined as

$$\frac{s_i^2 \hat{\sigma}_{t,i}^2}{\sigma^2 \log\left(1 + s_i^2 \frac{\hat{\sigma}_{t,i}^2}{\sigma^2}\right)} \leq \frac{\hat{\sigma}_{t,i}^2}{\sigma^2 \log\left(1 + \frac{\hat{\sigma}_{t,i}^2}{\sigma^2}\right)} \leq \frac{\sigma_{0,i}^2}{\sigma^2 \log\left(1 + \frac{\sigma_{0,i}^2}{\sigma^2}\right)} = c_i\,.$$

The first inequality holds because $s_i \leq 1$, and $ax/\log(1 + ax) \leq x/\log(1 + x)$ for any $a \in [0, 1]$ and $x > 0$. The second inequality holds because $x/\log(1 + x)$ is maximized when x is, which happens at $\hat{\sigma}_{t,i} = \sigma_{0,i}$.

For any $c \geq 1$ and node $i \in [L]$,

$$\log\left(1 + s_i^2 \frac{\hat{\sigma}_{t,i}^2}{\sigma^2}\right) = c^{L-i}c^{i-L}\log\left(1 + s_i^2 \frac{\hat{\sigma}_{t,i}^2}{\sigma^2}\right) \leq c^{L-i}\log\left(1 + c^{i-L}s_i^2 \frac{\hat{\sigma}_{t,i}^2}{\sigma^2}\right)\,,$$

where the inequality holds because $a\log(1 + x) \leq \log(1 + ax)$ for any $a \in [0, 1]$ and $x > 0$. Moreover,

$$\log\left(1 + c^{i-L}s_i^2 \frac{\hat{\sigma}_{t,i}^2}{\sigma^2}\right) = \log\left(\hat{\sigma}_{t,i}^{-2} + \frac{c^{i-L}s_i^2}{\sigma^2}\right) - \log(\hat{\sigma}_{t,i}^{-2}) \leq \log(\hat{\sigma}_{t+1,i}^{-2}) - \log(\hat{\sigma}_{t,i}^{-2})\,,$$

where the last step is by Lemma 3. Now we chain all inequalities, switch to the full notation, and get

$$\bar{\sigma}_{t,A_t}^2 \leq \sigma^2 \sum_{i=1}^{L_t} c^{L_t - i} c_{\psi_t(i)}[\log(\hat{\sigma}_{t+1,\psi_t(i)}^{-2}) - \log(\hat{\sigma}_{t,\psi_t(i)}^{-2})]\,.$$

Finally, we sum up the above upper bound over all rounds t. Let h_i be the maximum length of any path from node i to its descendant. Due to $c \geq 1$ and telescoping in the above decomposition, we get

$$\mathcal{V}(n) \leq \sigma^2 \sum_{i \in \mathcal{N}} c^{h_i} c_i [\log(\hat{\sigma}_{n+1,i}^{-2}) - \log(\sigma_{0,i}^{-2})] = \sigma^2 \sum_{i \in \mathcal{N}} c^{h_i} c_i \log(\sigma_{0,i}^2 \hat{\sigma}_{n+1,i}^{-2})\,.$$

For an action node i, $\hat{\sigma}_{n+1,i}^{-2} \leq \sigma_{0,i}^{-2} + \sigma^{-2}n$, and thus

$$\log(\sigma_{0,i}^2 \hat{\sigma}_{n+1,i}^{-2}) \leq \log\left(1 + \frac{\sigma_{0,i}^2 n}{\sigma^2}\right)\,.$$

For an internal node i, $\hat{\sigma}_{n+1,i}^{-2} \leq \sigma_{0,i}^{-2} + \sum_{j \in \mathsf{ch}(i)} \sigma_{0,j}^{-2}$, and thus

$$\log(\sigma_{0,i}^2 \hat{\sigma}_{n+1,i}^{-2}) \leq \log\left(1 + \sigma_{0,i}^2 \sum_{j \in \mathsf{ch}(i)} \sigma_{0,j}^{-2}\right) .$$

This completes the proof.

C. Contextual Regret Bound Proofs

This section contains proofs of our contextual regret bound and supporting lemmas.

C.1. Proof of Theorem 5

First, we apply Lemma 1 and get that

$$\mathcal{BR}(n) \leq \sqrt{2n\mathcal{V}(n)\log(1/\delta)} + \sqrt{2/\pi}\sigma_{\max}Kn\delta ,$$

where $\mathcal{V}(n) = \mathbb{E}\left[\sum_{t=1}^{n} \bar{\sigma}_{t,A_t}^2\right]$ and $\bar{\sigma}_{t,A_t}^2 = \|X_t\|_{\bar{\Sigma}_{t,A_t}}^2$ denotes the marginal posterior variance in node A_t in round t, in the direction of context X_t. We derive a worst-case upper bound on $\mathcal{V}(n)$ next.

We start with a worst-case upper in any round t. Since round t is fixed, we write L instead of L_t and refer to node $\psi_t(i)$ by i for any $i \in [L_t]$. Let $S_i = \prod_{j=i+1}^{L} \Sigma_{0,j}^{-1}\hat{\Sigma}_{t,j}$. Then by Lemma 8,

$$\bar{\sigma}_{t,A_t}^2 = X_t^\top \bar{\Sigma}_{t,A_t} X_t = \sigma^2 \sum_{i=1}^{L} \sigma^{-2} X_t^\top S_i^\top \hat{\Sigma}_{t,i} S_i X_t \leq \sigma^2 \sum_{i=1}^{L} c_i \log(1 + \sigma^{-2} X_t^\top S_i^\top \hat{\Sigma}_{t,i} S_i X_t) ,$$

where c_i is an upper bound at node i defined as

$$\frac{X_t^\top S_i^\top \hat{\Sigma}_{t,i} S_i X_t}{\sigma^2 \log(1 + \sigma^{-2} X_t^\top S_i^\top \hat{\Sigma}_{t,i} S_i X_t)} \leq \frac{\lambda_1(\hat{\Sigma}_{t,i})}{\sigma^2 \log(1 + \sigma^{-2}\lambda_d(\hat{\Sigma}_{t,i}))} \leq \frac{\sigma_{0,i}^2}{\sigma^2 \log\left(1 + \frac{\sigma_{0,i}^2}{\sigma^2}\right)} = c_i .$$

The first inequality holds because $\lambda_1(S_i) \leq 1$ and $\|X_t\|_2 \leq 1$, where the former follows from the definitions of $\hat{\Sigma}_{t,j}$ and $\Sigma_{0,j} = \sigma_{0,j}^2 I_d$. We also use the fact that $ax/\log(1 + ax) \leq x/\log(1 + x)$ holds for any $a \in [0, 1]$ and $x > 0$. The second inequality holds because $\lambda_1(\hat{\Sigma}_{t,i}) \leq \lambda_1(\Sigma_{0,i}) = \sigma_{0,i}^2$. We also use that $x/\log(1 + x)$ is maximized when x is.

For any $c \geq 1$ and node $i \in [L]$,

$$\begin{aligned}\log(1 + \sigma^{-2} X_t^\top S_i^\top \hat{\Sigma}_{t,i} S_i X_t) &= c^{L-i} c^{i-L} \log(1 + \sigma^{-2} X_t^\top S_i^\top \hat{\Sigma}_{t,i} S_i X_t) \\ &\leq c^{L-i} \log(1 + \sigma^{-2} c^{i-L} X_t^\top S_i^\top \hat{\Sigma}_{t,i} S_i X_t) ,\end{aligned}$$

where the inequality holds because $a\log(1 + x) \leq \log(1 + ax)$ for any $a \in [0, 1]$ and $x > 0$. Moreover,

$$\begin{aligned}\log(1 + \sigma^{-2} c^{i-L} X_t^\top S_i^\top \hat{\Sigma}_{t,i} S_i X_t) &= \log\det(I_d + \sigma^{-2} c^{i-L} \hat{\Sigma}_{t,i}^{\frac{1}{2}} S_i X_t X_t^\top S_i^\top \hat{\Sigma}_{t,i}^{\frac{1}{2}}) \\ &= \log\det(\hat{\Sigma}_{t,i}^{-1} + \sigma^{-2} c^{i-L} S_i X_t X_t^\top S_i^\top) - \log\det(\hat{\Sigma}_{t,i}^{-1}) \\ &\leq \log\det(\hat{\Sigma}_{t+1,i}^{-1}) - \log(\hat{\Sigma}_{t,i}^{-1}) ,\end{aligned}$$

where the last step is by Lemma 9. Now we chain all inequalities, switch to the full notation, and get

$$\bar{\sigma}_{t,A_t}^2 \leq \sigma^2 \sum_{i=1}^{L_t} c^{L_t - i} c_{\psi_t(i)} [\log\det(\hat{\Sigma}_{t+1,\psi_t(i)}^{-1}) - \log\det(\hat{\Sigma}_{t,\psi_t(i)}^{-1})] .$$

Finally, we sum up the above upper bound over all rounds t. Let h_i be the maximum length of any path from node i to its descendant. Due to $c \geq 1$ and telescoping in the above decomposition, we get

$$\mathcal{V}(n) \leq \sigma^2 \sum_{i \in \mathcal{N}} c^{h_i} c_i [\log\det(\hat{\Sigma}_{n+1,i}^{-1}) - \log\det(\Sigma_{0,i}^{-1})] = \sigma^2 \sum_{i \in \mathcal{N}} c^{h_i} c_i \log\det(\Sigma_{0,i}^{\frac{1}{2}} \hat{\Sigma}_{n+1,i}^{-1} \Sigma_{0,i}^{\frac{1}{2}}) .$$

For an action node i,

$$\log\det(\Sigma_{0,i}^{\frac{1}{2}}\hat{\Sigma}_{n+1,i}^{-1}\Sigma_{0,i}^{\frac{1}{2}}) \leq d\log\left(\frac{1}{d}\operatorname{tr}(\Sigma_{0,i}^{\frac{1}{2}}\hat{\Sigma}_{n+1,i}^{-1}\Sigma_{0,i}^{\frac{1}{2}})\right) \leq d\log\left(1 + \frac{\sigma_{0,i}^2 n}{\sigma^2 d}\right).$$

For an internal node i,

$$\log\det(\Sigma_{0,i}^{\frac{1}{2}}\hat{\Sigma}_{n+1,i}^{-1}\Sigma_{0,i}^{\frac{1}{2}}) \leq d\log\left(\frac{1}{d}\operatorname{tr}(\Sigma_{0,i}^{\frac{1}{2}}\hat{\Sigma}_{n+1,i}^{-1}\Sigma_{0,i}^{\frac{1}{2}})\right) \leq d\log\left(1 + \sigma_{0,i}^2 \sum_{j \in \mathsf{ch}(i)} \sigma_{0,j}^{-2}\right).$$

This completes the proof.

C.2. Supporting Lemmas

The first claim is a matrix generalization of Lemma 2.

Lemma 8. *In any round t, the marginal posterior covariance in action node A_t decomposes as*

$$\bar{\Sigma}_{t,A_t} = \sum_{i=1}^{L_t} S_i^\top \hat{\Sigma}_{t,\psi_t(i)} S_i\,,$$

where $S_i = \prod_{j=i+1}^{L_t} \Sigma_{0,\psi_t(j)}^{-1}\hat{\Sigma}_{t,\psi_t(j)}$.

Proof. The proof is exactly the same as in Appendix B.2, except that the variances are replaced by covariances. □

The second claim is a matrix generalization of Lemma 3.

Lemma 9. *Fix any round t and $i \in [L_t]$. Then*

$$\hat{\Sigma}_{t+1,\psi_t(i)}^{-1} - \hat{\Sigma}_{t,\psi_t(i)}^{-1} \succeq \sigma^{-2} c^{i-L_t} S_i X_t X_t^\top S_i^\top\,,$$

where $c = 1 + \sigma_{0,\max}^2/\sigma^2$ and S_i is defined in Lemma 8.

Proof. Since round t is fixed, we write L instead of L_t and refer to node $\psi_t(i)$ by i for any $i \in [L_t]$.

For an action node i, the claim holds trivially since $\hat{\Sigma}_{t+1,i}^{-1} - \hat{\Sigma}_{t,i}^{-1} = \sigma^{-2}X_tX_t^\top$. The rest of the proof is for internal nodes. We start with proving that for any node i, $\tilde{\Sigma}_{t+1,i}^{-1} - \tilde{\Sigma}_{t,i}^{-1} = vv^\top$ is a rank-1 matrix for some $v \in \mathbb{R}^d$, such that

$$vv^\top \succeq \sigma^{-2} c^{i-L-1} S_{i-1} X_t X_t^\top S_{i-1}\,, \quad v^\top v \leq \sigma^{-2}\,. \tag{16}$$

The proof is by induction.

The basis of the induction is that the claim holds for action nodes. Let node $i = L$ be an action node. We apply (11) and get

$$\begin{aligned}
\tilde{\Sigma}_{t+1,i}^{-1} - \tilde{\Sigma}_{t,i}^{-1} &= (\Sigma_{0,i} + (G_{t,i} + \sigma^{-2}X_tX_t^\top)^{-1})^{-1} - (\Sigma_{0,i} + G_{t,i}^{-1})^{-1} \\
&= \Sigma_{0,i}^{-1}[(\Sigma_{0,i}^{-1} + G_{t,i})^{-1} - (\Sigma_{0,i}^{-1} + G_{t,i} + \sigma^{-2}X_tX_t^\top)^{-1}]\Sigma_{0,i}^{-1} \\
&= \Sigma_{0,i}^{-1}(\Sigma_{0,i}^{-1} + G_{t,i})^{-1}\frac{\sigma^{-2}X_tX_t^\top}{1 + \sigma^{-2}X_t^\top(\Sigma_{0,i}^{-1} + G_{t,i})^{-1}X_t}(\Sigma_{0,i}^{-1} + G_{t,i})^{-1}\Sigma_{0,i}^{-1} \\
&= \Sigma_{0,i}^{-1}\hat{\Sigma}_{t,i}\frac{\sigma^{-2}X_tX_t^\top}{1 + \sigma^{-2}X_t^\top\hat{\Sigma}_{t,i}X_t}\hat{\Sigma}_{t,i}\Sigma_{0,i}^{-1}\,.
\end{aligned}$$

The second and third equalities are by the Woodbury and Sherman-Morrison formulas, respectively. The matrix is rank 1. The lower bound follows from $\Sigma_{0,i} = \sigma_{0,i}^2 I_d$ and $\|X_t\|_2 \leq 1$, which can be used to derive

$$1 + \sigma^{-2}X_t^\top\hat{\Sigma}_{t,i}X_t \leq 1 + \sigma^{-2}\lambda_1(\hat{\Sigma}_{t,i}) \leq 1 + \sigma^{-2}\sigma_{0,i}^2 \leq c\,.$$

The upper bound uses that $\|\Sigma_{0,i}^{-1}\hat{\Sigma}_{t,i} X_t\|_2 \leq 1$ and $1 + x \geq 1$ for any $x \geq 0$.

In the inductive step, we assume that (16) holds for node $i + 1$ and prove it for node i. Note that node $i + 1$ is the only child of node i where the posterior between rounds t and $t + 1$ changes. Let $M = \sum_{j \in \mathsf{ch}(i)} \tilde{\Sigma}_{t,j}^{-1}$ and $\varepsilon = \tilde{\Sigma}_{t+1,i+1}^{-1} - \tilde{\Sigma}_{t,i+1}^{-1}$. Now we apply (12) and get

$$\begin{aligned}
\tilde{\Sigma}_{t+1,i}^{-1} - \tilde{\Sigma}_{t,i}^{-1} &= (\Sigma_{0,i} + (M + \varepsilon)^{-1})^{-1} - (\Sigma_{0,i} + M^{-1})^{-1} \\
&= \Sigma_{0,i}^{-1}[(\Sigma_{0,i}^{-1} + M)^{-1} - (\Sigma_{0,i}^{-1} + M + \varepsilon)^{-1}]\Sigma_{0,i}^{-1} \\
&= \Sigma_{0,i}^{-1}(\Sigma_{0,i}^{-1} + M)^{-1} \frac{\varepsilon}{1 + v^\top (\Sigma_{0,i}^{-1} + M)^{-1} v} (\Sigma_{0,i}^{-1} + M)^{-1} \Sigma_{0,i}^{-1} \\
&= \Sigma_{0,i}^{-1} \hat{\Sigma}_{t,i} \frac{v v^\top}{1 + v^\top \hat{\Sigma}_{t,i} v} \hat{\Sigma}_{t,i} \Sigma_{0,i}^{-1} \,,
\end{aligned}$$

where $\varepsilon = v v^\top$. The matrix is rank 1, because by the inductive argument ε is rank 1. As in the proof for the action node, the second and third equalities are by the Woodbury and Sherman-Morrison formulas. The lower bound is derived using

$$1 + v^\top \hat{\Sigma}_{t,i} v \leq 1 + \|v\|_2^2 \lambda_1(\hat{\Sigma}_{t,i}) \leq 1 + \sigma^{-2} \sigma_{0,i}^2 \leq c \,.$$

This follows from $\Sigma_{0,i} = \sigma_{0,i}^2 I_d$ and $v^\top v \leq \sigma^{-2}$, where the latter is by the inductive argument. The upper bound uses that $\|\Sigma_{0,i}^{-1} \hat{\Sigma}_{t,i} v\|_2 \leq \sigma^{-1}$ and $1 + x \geq 1$ for any $x \geq 0$.

To complete the proof, we have from (10) that

$$\hat{\Sigma}_{t+1,i}^{-1} - \hat{\Sigma}_{t,i}^{-1} = \tilde{\Sigma}_{t+1,i+1}^{-1} - \tilde{\Sigma}_{t,i+1}^{-1}$$

holds for any internal node i. □